Practical Assessment, Research, and Evaluation

Volume 9 *Volume 9, 2004* Article 1

2004

Effects of Handwriting and Computer-Print on Composition Scores: A Follow-up to Powers, Fowles, Farnum, & Ramsey

Russell Michael

Wei Tao

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 9, Number 1, January, 2004 ISSN=1531-7714

Effects of Handwriting and Computer-Print on Composition Scores: A Follow-up to Powers, Fowles, Farnum, & Ramsey.

Michael Russell & Wei Tao
Boston College

Powers, Fowles, Farnum, & Ramsey (1994) report that essays presented in handwritten form receive higher scores than the same essays presented as computer-text. The study presented here partially replicates the study conducted by Powers et al. In addition, this study explores the extent to which several types of errors in student writing are more visible when essays are presented as computer-text. The two experiments presented here partially confirm previous findings that essays presented in computer print receive lower scores and that part of this effect may result from the increased visibility of errors in computer-text. Through interviews with readers, this study also suggests that the higher standards and expectations readers have for text presented as computer print and the ability of some readers to identify with students and see their effort when they handwrite essays may impact the scores they award. Unlike Powers et al, however, this study does not find evidence that altering the spacing of essays presented in computer print in order to increase the perceived length consistently reduces the presentation effect.

As the Educational Testing Service began to offer The Praxis Series: Professional Assessments for Beginning Teachers on computer, Powers, Fowles, Farnum and Ramsey (1994) undertook a study that examined the equivalence of scores awarded to responses presented as computer-printed text or in handwritten form. Prior to the experiment conducted by Powers et al. (1994), several studies focused on the influence that "neat" versus "sloppy" penmanship had on scores awarded to essays. This body of research consistently reports that essays presented with neater penmanship receive higher scores than those presented with sloppy penmanship (Chase, 1986; Marshall & Powers, 1969; Markham, 1976; Bull & Stevens, 1979). Thus, one would expect that essays presented as neatly formatted computer-printed text would receive higher scores than essays presented in handwritten form. Surprisingly, Powers et al. (1994) found the exact opposite: Raters awarded higher scores to responses presented in handwritten form as compared to the exact same responses presented as computer-printed text.

To explain this seemingly contradictory finding, Powers et al. (1994) offered several hypotheses, some of which drew upon their work as well as the work of Arnold, Legas, Obler, Zpacheco, Russell, and Umbdenstock (as summarized by Powers et al., 1994). These hypotheses included:

A. Readers may have expected fully edited and polished final products when presented as computer-printed text and thus had higher expectations for these essays;
B. Handwritten text caused the reader to feel closer to the writer which "allowed for a closer identification of the writer's individual voice as a strong and important aspect of the essay" (as quoted in Powers et al.);
C. Readers may have given handwritten responses the benefit of the doubt when they encountered sloppy or hard-to-read text;
D. Hand-written responses appeared longer and thus appeared to have been the result of greater effort.

To examine the final hypothesis, Powers et al. (1994) conducted a small follow-up study during which computer-printed responses were double-spaced to make them appear longer. During this follow-up study, training procedures were also modified such that readers were informed of the presentation effect and were instructed to apply the same criteria to handwritten and computer-printed responses. The combination of supplemental training and double-spacing of computer-printed responses reduced the presentation effect, but did not eliminate it.

Although the study conducted by Powers et al. (1994) was published nearly ten years ago, no further work has been performed to examine the presentation effect. During this same time period, however, the use of computers for writing in schools has increased rapidly. Concurrently, the increased reliance on tests to make high-stakes decisions about student and school performance, have sparked calls for testing programs to allow students the option of composing responses on paper or with a computer. These calls are bolstered by the results of a series of studies that examined the effects of different composition modes on students' writing performance and which found that the achievement of

students accustomed to writing with a computer are underestimated by hand-written open-ended tests (Russell & Haney, 1997; Russell, 1999; Russell & Plati, 2000). In all three studies, students were randomly assigned to perform open-ended or essay items on computer or on paper. And in each study, the difference in performance on paper and on computer ranged from an effect size (d) of about .4 to just over 1.0, with students who were accustomed to writing on computers performing better when they were able to produce responses to an open-ended item using a computer.

In practical terms, the mode of administration found in the first study indicated that when students accustomed to writing on computer were forced to use paper-and-pencil, only 30 percent of students performed at a "passing" level; when they wrote on computer, 67 percent "passed." In a second study, the difference in performance on paper versus on computer for students who could keyboard approximately 20 words a minute was larger than the amount students' scores typically change between grade 7 and grade 8 on standardized tests. However, for students who were not accustomed to writing on computer and could only keyboard at relatively low levels, taking the tests on computer diminished performance. Finally, a third study that focused on the Massachusetts Comprehensive Assessment Systems Language Arts Tests demonstrated that removing the mode of administration effect for writing items would have a dramatic impact on the study district's results. Based on 1999 MCAS results, the study estimated that 19% of the fourth graders classified as "Needs Improvement" would move up to the "Proficient" performance level. An additional 5% of students who were classified as "Proficient" would be deemed "Advanced".

The Powers experiment and the 3 studies conducted by Russell and his colleagues indicate that two variables may influence students' writing scores – mode of composition and mode of presentation. "Mode of composition" refers to how students produce their essays – keyboard composition or paper/pencil composition. "Mode of presentation" refers to essay formats presented to the readers as handwritten or as computer text. To limit the influence of mode of composition on student performance, Russell and Haney (2000) suggest that students be allowed to select the mode in which open-ended responses are composed. Although no state testing programs in the United States have embraced this option, the Province of Alberta has employed this strategy for its graduation testing program for the past decade and have seen the percentage of students opting to compose responses with a computer increase from 6.7% in 1996 to 24.5% in 2000 (A. Sakyi, personal communication, April 26, 2000). However, while allowing students to choose their mode of composition may decrease the mode of composition effect, it introduces the mode of presentation effect.

The study presented below was undertaken to replicate Powers et al. (1994) work on the mode of presentation and to further probe the causes of the presentation effect.

Methodology

Two sets of analyses were undertaken to examine the presentation effect first reported by Powers et al. The first set of analyses focuses on the presence or absence of a presentation effect for essays produced by students in grades four, eight and ten. The second set of analyses was undertaken to identify factors that may be influencing raters' scores and thus be causing the presentation effect. The methodology for both sets of analyses are described separately below.

Experiment 1: Presentation Effect Partial Replication Experiment

As part of a larger study that focused on the mode of administration effect on the MCAS Composition items, Russell and Plati (2000) transcribed approximately 240 hand-written responses to computer-text. These essays were produced by students in grades 4, 8 and 12 and were in response to grade specific items that appeared on the 1999 Massachusetts Comprehensive Assessment System (MCAS) Language Arts Test. The responses were produced by students in a suburban district outside of Boston that tends to perform well on the state's tests.

For the analyses reported here, sixty student essays were randomly selected from grade 8 and 10 respectively and all fifty-two essays available in grade 4 were selected. In all three grade levels, the essays were originally produced on paper and were then transcribed verbatim (including all spelling, grammar, and punctuation errors) into computer format by the research team. Within each grade level, all responses were presented to raters in three ways:

1. handwritten;
2. single spaced twelve point Times Roman computer-text, and;
3. double-spaced fourteen point Times Roman computer-text.

Thus, the purpose of the analyses was to examine the extent to which raters awarded the same scores to the same responses presented in three different formats.

To ensure the precision in transcription, the following procedures were adopted. When transcribing responses from their original handwritten form to computer text, responses were first transcribed verbatim into the computer. The transcriber then printed out the computer version and compared it word by word with the original, making corrections as needed. A second person then compared these corrected transcriptions with the originals and made additional changes as needed. Following this process, a sample of 10 responses was checked a third time. Out of 3,524 words of text, only three errors were found and in two cases a word that had been misspelled in the original was spelled correctly in the transcribed text. Thus, while slight differences may exist between the original handwritten and the transcribed

versions, these differences are likely to have a very minor effect on rater's scores.

To control for any differences in the accuracy with which raters apply the scoring criteria, a counterbalanced design was used in which raters scored twenty of each presentation format. Six raters were employed for each grade level. With the exception of one rater, who was a graduate student, all raters were classroom teachers who taught English/Language Arts at the same grade level. Prior to participating in the study, all raters were informed that raters were needed to score a sample of responses completed by students in a local school district in preparation for the upcoming state writing test. It should be noted that the same criteria used to select raters for the State language arts test were applied when recruiting raters for this study. The assignment of raters' to essays is shown in table 1. As shown, different presentation formats of the same essays were scored by different pairs of raters. For example, the handwritten format of essay No. 1-20 were scored by rater 1 and 2, single-space computer text of the same essays were scored by rater 3 and 4, and the double-space computer text were scored by rater 5 and 6. In addition, note that all pairs of raters scored essays presented in each format, but none of the raters scored the same essay twice.

Table 1: Essay Distribution Among the 6 Raters

Essays	Essay Format		
	Handwritten	Single-space computer text	Double-space computer text
#1 ~ 20	Rater 1, 2	Rater 3,4	Rater 5, 6
#21 ~ 40	Rater 5, 6	Rater 1, 2	Rater 3,4
#41 ~ 60	Rater 3,4	Rater 5, 6	Rater 1, 2

Following MCAS scoring procedures, all responses presented in a given format were double-scored and scores awarded by each rater were aggregated into a single score by adding the two scores.

For all of the items, the scoring criteria developed for MCAS were used (Massachusetts Department of Education, 2000a). The MCAS scoring guidelines for the composition items focused on two areas of writing, namely Topic/Idea Development and Standard English Conventions. The scale for Topic Development ranged from 1 to 6 and the scale for English Conventions ranged from 1 to 4, with one representing the lowest level of performance for both scales. Table 2 presents the category descriptions for each point on the two scales.

Table 2: Category Descriptions for MCAS Composition Rubrics

Score	Topic Development	English Standards
1	Little topic/idea development, organization, and/or details Little or no awareness of audience and/or task	Errors seriously interfere with communication AND Little control of sentence structure, grammar and usage and mechanics
2	Limited or weak topic/idea development, organization, and/or details Limited awareness of audience and/or task	Errors interfere somewhat with communication and/or Too many errors relative to the length of the essay or complexity of sentence structure, grammar and usage, and mechanics
3	Rudimentary topic/idea development and/or organization Basic supporting details Simplistic language	Errors do not interfere with communication and/or Few errors relative to length of essay or complexity of sentence structure, grammar and usage, and mechanics
4	Moderate topic/idea	Control of sentence structure, grammar and usage, and

	development and organization Adequate, relevant details Some variety in language	mechanics (length and complexity of essay provide opportunity for students to show control of standard English conventions)
5	Full topic/idea development Logical organization Strong details Appropriate use of language	
6	Rich topic/idea development Careful and/or subtle organization Effective/rich use of language	

In addition to the general descriptions, MCAS also provides anchor papers and benchmark papers presented in handwritten form for each category. These anchor and benchmark papers provide concrete examples of each performance level. The anchor and benchmark papers were first introduced to raters during the common scoring training session and were available to raters throughout the scoring process.

Following procedures used during the actual MCAS testing, inter-rater reliability was examined by comparing the percent agreement within one point between the two raters. As Table 3 indicates, agreement within one point was greater than 90% in all cases and was comparable to the agreement reported during actual MCAS scoring procedures. Exact agreement, however, was noticeably lower and correlation coefficients were moderate. It should be noted that mean scores were above the mid-point of each scale across all grade levels. For grades 4 and 10, raters employed the full score range when scoring student responses. However, for both Topic Development and English Standards, none of the raters used the lowest score (1). It should also be noted that for grade 10, there was also a moderate (-.40) negative skew to the distribution of scores. All other distributions were approximately normally distributed. Unfortunately, the MCAS Technical Report does not report percent exact agreement, correlation coefficients or information about the distribution of scores for these two scales, so it is not possible to fully compare the reliability found during this study with that of the actual MCAS composition scoring (Massachusetts Department of Education, 1999; 2000b).

Table 3: Percent Agreement between Raters

	Exact	Within One Point	Correlation
Grade 4			
Topic Development	44%	91%	.64
English Standards	54%	99%	.55
Grade 8			
Topic Development	46%	94%	.55
English Standards	52%	99%	.54

Grade 10			
Topic Development	54%	91%	.78
English Standards	71%	98%	.71

Experiment 2: Identifying Causes of Presentation Effect

As described above, Powers et al. (1994) identified several factors that may contribute to the presentation effect. One of these factors related to the visibility of mechanical and structural errors when responses are presented in computer-print format. To examine the extent to which mechanical and structural errors are more or less visible when responses are presented as computer-print or handwritten form, two readers were asked to read a sample of thirty grade eight responses. As they read each response, the readers were asked to mark the following types of errors:

- Spelling
- Punctuation
- Capitalization
- Awkward transitions
- Confusing phrases or sentences

To reduce reader bias, each reader read one set of fifteen responses presented in handwritten form and a second set presented in computer print form. The responses were intermingled and counterbalanced such that each reader read each response only one time. Also note that the readers were asked to mark all errors they encountered as they read through the response only one time.

After the readers were finished reading through the responses and marking all visible errors, each type of error was summed for each essay.

Findings

Experiment 1: Presentation Effect

Table 4 presents the mean score for each mode of presentation by grade level. In all grade levels, the mean score for responses presented in handwriting was higher than mean scores for responses presented as computer-text. In grade eight, the presentation effect resulted in more than a two-point difference between scores awarded to handwritten responses versus computer printed responses. In grade four and eight, there was little difference between scores awarded to single spaced and double spaced responses. In grade ten, however, double-spaced responses tended to be awarded lower scores than single-spaced responses.

Table 4: Descriptive Statistics for Mode of Presentation and ANOVA test on Group Means

	Descriptive Statistics			ANOVA test on group means			
	Handwritten	Single Spaced	Double Spaced				
	Mean (Std. Dev.)	Mean (Std. Dev.)	Mean (Std. Dev.)	Between	Within	F	Sig.
Grade 4 (N=52)							
Topic Development*	8.3 (1.71)	7.0 (1.99)	7.4 (2.06)	23.6	3.7	6.4	**.002**
English	6.5	5.8	5.7	9.9	1.6	6.2	**.002**

Standards*+	(1.11)	(1.26)	(1.38)				
Total Score*+	14.8 (2.57)	12.8 (3.04)	13.1 (3.29)	59.5	8.9	6.7	.002
Grade 8 (N=60)							
Topic Development*+	9.2 (1.69)	7.9 (1.67)	7.9 (1.78)	33.4	2.9	11.3	<.001
English Standards*+	6.9 (0.97)	6.0 (1.07)	6.1 (1.17)	14.9	1.2	13.0	<.001
Total Score*+	16.1 (2.51)	13.9 (2.50)	14.0 (2.71)	92.7	6.6	14.0	<.001
Grade 10 (N=60)							
Topic Development+	8.6 (2.33)	7.7 (2.32)	6.9 (2.22)	35.3	5.2	6.8	.001
English Standards*+	6.8 (1.27)	6.2 (1.33)	6.0 (1.37)	10.7	1.8	5.8	.003
Total Score*+	15.3 (3.40)	13.9 (3.24)	13.0 (3.23)	82.4	10.8	7.6	<.001

Tukey HSD post-hoc comparisons were conducted to examine the statistical significance of differences between each item format.
* indicates a statistically significant difference at the .05 level between Handwritten and Single Spaced formats.
+ indicates a statistically significant difference at the .o5 level between Handwritten and Double Spaced formats.

To examine whether the mean differences in sub-scaled scores and total scores were statistically significant, a one-way analysis of variance was performed within each grade level. As Table 4 indicates, the differences among the three modes of presentations were significant within each grade level. To compare sub-scale and total score means for each mode of presentation, Tukey HSD post-hoc comparisons were performed. Table 4 also indicates that for grades four, eight and ten, differences between both single-spaced and handwritten responses and between double-spaced and handwritten responses were statistically significant. Differences between single and double-spaced responses were not statistically significant. In other words, single-spaced computer text and double-spaced computer text are homogeneous sub-set, or similar with each other; while both of them are significantly differently from handwritten text.

Experiment 2: Identifying Causes of Presentation Effect

As described above, a second set of analyses was performed to identify the extent to which errors are more or less visible when responses are presented in handwritten or computer-printed form. Table 5 presents summary statistics for the five types of errors examined, namely spelling, punctuation, capitalization, awkward transitions, and confusing passages. With the exception of awkward transitions, readers identified more errors when the passages were presented in computer-printed form than when the same responses were presented in handwritten form. Moreover, the differences between detection of spelling errors, capitalization errors and confusing passages were all statistically significant.

Note that statistical significance for the t-tests reported in Table 5 was not adjusted to account for multiple comparisons. Given that five comparisons were made for each group, there is an increased probability that reported differences

occurred by chance. Employing the Dunn approach to multiple comparisons (see Glass & Hopkins, 1984), a for *c* multiple comparisons, a_{pc}, is related to simple a for a single comparison as follows:

$$a_{pc} = 1 - (1\text{-}a)^{1/c}$$

Hence, for five comparisons the adjusted value of a simple 0.05 alpha level becomes 0.01. Analogously, a simple alpha level of 0.01 for a simple comparison becomes 0.002. After adjusting for multiple comparisons, differences in the visibility of spelling errors and confusing passages remain statistically significant.

Table 5: Summary Statistics and T-test for Observed Errors

	Handwritten Form N=30		Computer-Print Form N=30			
Error Type	**Mean**	**St. Dev.**	**Mean**	**St. Dev.**	**T-Statistic**	**Sig.**
Spelling	4.69	4.58	8.23	5.2	2.87	.006
Punctuation	.84	1.17	1.32	1.47	1.44	.160
Capitalization	.22	.49	1.26	2.63	2.20	.030
Awkward Transition	3.47	2.23	2.97	1.99	.94	.350
Confusing Passage	.25	.72	1.06	1.03	3.65	.001

Discussion

The results presented above partially confirm the findings reported by Powers et al. (1994). In all three grade levels, responses presented in handwritten form received significantly higher scores than did the exact same responses presented in computer print form. Unlike Powers et al., however, the analyses reported here found that altering the formatting of computer-printed responses to make them appear longer did not have a significant impact on raters' scores. In grade four, scores awarded to double-spaced computer-printed responses did receive slightly higher scores. But the opposite occurred in grade 10. This lack of confirmation, however, suggests that it may have been the supplemental training and not the double-spacing that reduced the presentation effect in Powers et al.'s study.

To further probe the factors that may cause the presentation effect, all of the raters were interviewed once they had completed scoring all sixty responses. During these interviews, raters were first asked if they found it easier or harder to score handwritten responses. Raters were then asked whether they found themselves applying different standards or criteria to handwritten responses as compared to the computer-printed responses. Finally, raters were told about the presentation effect found by Powers et al. and were asked what factors they think might contribute to the presentation effect.

Unanimously, raters agreed that it was easier to score the computer-printed responses simply because they were easier to read. All but two of the raters also felt that the double-spaced responses were easier to read than the single-spaced responses. One dissenting rater preferred the single-spaced responses because she could see more of the text and, in most cases, the response fit on a single page eliminating the need to flip to a second page.

Although none of the raters initially reported that they were applying different standards, once they were informed about the presentation effect (after they had completed scoring responses), raters offered several reasons why this effect may occur. All but four of the raters mentioned that they noticed many more mechanical errors in the computer printed responses. This observation was confirmed by the second set of analyses reported above in which spelling and capitalization errors were more visible when responses were presented with computer-text as compared to handwritten form. Three of these raters added that they had a hard time resisting correcting students mistakes as they read responses in computer printed form.

Five of the readers also mentioned that as they read the computer printed forms, they had to keep reminding themselves that these were not final drafts, but works in progress that had been produced using a computer. (Note that all raters were blind to the study and were told that the responses were produced during a two-hour block of time and

that students worked either on paper or on computer.)

Four raters also stated that they felt the handwritten essays were more personable and that they felt a stronger connection to the writer because of their handwriting. One of these raters added that the cross-outs and last minute changes created the sense that students who produced responses by hand "really tried hard," whereas "there was no clear evidence that students put in lots or no effort on computer." Finally, nearly all of the teachers stated that one factor that may cause the presentation effect relates to the fact that students in their classes generally create rough drafts on paper and then produce final drafts on computer. As a result, they are accustomed to thinking of computer-printed responses as final drafts. Moreover, most final drafts are submitted in double-space form. Although they did not initially feel that they applied different criteria, in hindsight they all felt that may have been "harsher" on the computer printed responses because they thought of them as final drafts.

Given the current policy in Alberta and increased calls in the United States to provide students with the choice of composing essay tests on paper or with a computer, this study provides further evidence that essays presented in hand-written form may be scored more leniently than essays presented as computer-printed text. While the mode of administration reported by Russell and his colleagues suggests that students accustomed to writing on computer are put at a disadvantage when forced to perform essay tests on paper, Powers et al. and the study presented here suggests that these same students may also be put at a disadvantage when it comes to scoring if allowed to compose their essays on computer. While one solution might be to transcribe all handwritten responses to computer-text, this strategy is infeasible for large testing programs. Instead, efforts should focus on developing strategies to "train away" the presentation effect and/or to statistically adjust scores.

References

Alberta Learning. (2000). *Directions for Administration, Administrators Manual,* Diploma Examination Program.

Bull, R. & Stevens, J. (1979). The effects of attractiveness of writer and penmanship on essay grades. *Journal of Occupational Psychology*, 52, 53-59.

Chase, C. I. (1986). Essay test scoring: Interaction of relevant variables. *Journal of Educational Measurement*, 23(1), 33-41.

Glass, G. & Hopkins, K. (1984). *Statistical Methods in Education and Psychology*. Boston, MA: Allyn and Bacon.

Markham, L. R. (1976). Influences of handwriting quality on teacher evaluation of written work. *American Educational Research Journal*, 13(4), 277-283.

Marshall, J. C. & Powers, J. C. (1969). Writing neatness, composition errors, and essay grades. *Journal of Educational Measurement*, 6, 97-101.

Massachusetts Department of Education. (1999). *Massachusetts Comprehensive Assessment System: 1998 Technical Report*. Malden, MA.

Massachusetts Department of Education. (2000a). 1999 MCAS Sample Student Work and Scoring Guides. Available online: http://www.doe.mass.edu/mcas/student/1999/.

Massachusetts Department of Education. (2000b). *1999 MCAS Technical Report*. Malden, MA.

Powers, D., Fowles, M, Farnum, M, & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement*, 31(3), 220-233.

Russell, M. & Haney, W. (1997). Testing writing on computers: an experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives*, 5(3), Available online: http://olam.ed.asu.edu/epaa/v5n3.html.

Russell, M. & Haney, W. (2000). Bridging the Gap Between Testing and Technology in Schools. *Education Policy Analysis Archives*, 8(19), Available online: http://epaa.asu.edu/epaa/v8n19.html.

Russell, M. & Plati, T. (2000). Mode of Administration Effects on MCAS Composition Performance for Grades Four, Eight and Ten. A report submitted to the Massachusetts Department of Education by the National Board on Educational Testing and Public Policy, Available online: http://nbetpp.bc.edu/reports.html.

Russell, M. (1999). Testing Writing on Computers: A Follow-up Study Comparing Performance on Computer and on Paper. *Educational Policy Analysis Archives*, 7(20). Available online: http://epaa.asu.edu/epaa/v7n20/.

Authors

Michael Russell
Technology and Assessment Study Collaborative
Center for the Study of Testing, Evaluation and Educational Policy
Boston College
323 Campion Hall
Chestnut Hill, MA 02481

617 552-0889
russelmh@bc.edu

Wei Tao
Lynch School of Education
Boston College
332G Campion Hall
Chestnut Hill, MA 02467

617-552-3574
taowe@bc.edu

Descriptors: Essay Tests; Test Construction; Writing Prompts; Computer Uses in Education; * Essays; * Handwriting; Interrater Reliability; * Scoring; * Word Processing; * Writing Evaluation

Citation: Michael Russell & Wei Tao (2004). Effects of handwriting and computer-print on composition scores: a follow-up to powers, fowles, farnum, & ramsey. *Practical Assessment, Research & Evaluation*, 9(1). Available online: http://PAREonline.net/getvn.asp?v=9&n=1.